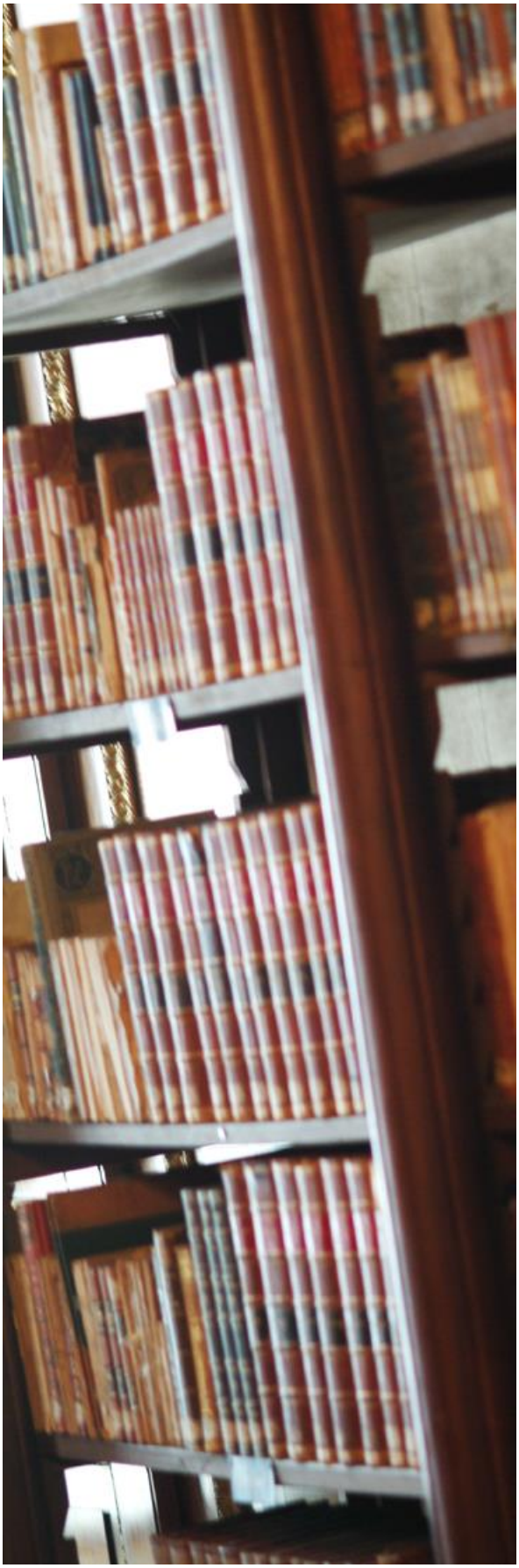


Internet Archive Book Digitization Process

(Note: This document is for information purposes only. Processes and workflow may change at any time, with or without notice. Internet Archive is not obligated to support the items in this document.)

April 23, 2015 – Rev. Z

Prepared by the Internet Archive Digital Libraries Team

Table of Contents

1. Introduction to Internet Archive

A. Background

Before settling upon the current operations workflow, engineers at Internet Archive (IA) tested and evaluated several commercially available book-digitizing devices. After carefully reviewing the condition, variety and library requirements for materials to be digitized, it was decided that developing an in-house digitizing process, complete with equipment and software, would provide an optimal balance of quality and efficiency while also ensuring that the materials being digitized would not be damaged. The IA-built equipment, software and process were reviewed with library preservation experts, field-tested, and has subsequently been used in IA-run Digitization Centers around the world. To date, over 400 million pages have been digitized using IA's non-destructive digitization method.

Periodically, new digitization equipment is reviewed to ensure that the most optimal software and hardware is in use. The workflow is also reviewed frequently in order to incorporate the discovery of new "best practices". All Digitization Centers follow the same general workflow described in this document, with any minor deviations being based upon specific library requirements. With only a few exceptions, Digitization Centers are managed and staffed by trained IA personnel.

B. Non-Destructive Digitization Station

The Scribe workstation is comprised of a sturdy aluminum frame that supports two adjustable camera mounting rails, two color cameras that capture both recto and verso pages of each book, a floating V-shaped book-cradle whose angled design minimizes stress placed on materials, a glass platen that is raised and lowered by means of a foot pedal, two banks of museum grade lights that illuminate the book, and one computer that captures the color images from the camera and performs some of the pre-processing. Once the book is digitized and an on-site Quality Assurance process is completed, the captured images are uploaded via RSYNC to processing computers located in California.

Book Cradle

Book Cradle and Foot Pedal Which Raises and Lowers the Glass Platen

Scribe Digitization Station

2. Technical Details and Specifications

A. Mechanical, Electrical, and Gross Size Dimensions

i. 1 ampere per Scribe, 144 watts of heat generated per Scribe (standard UK/USA voltage).

ii. Approximately 100 cubic feet of work area per Scribe.

iii. Scribe footprint when installed: 68" long x 37" wide x 79" high (172cm x 93cm x 200cm).

iv. Recommended door width – 34.5" (87.6 cm)

v. Width dimensions for delivery through doorways/into work area (when crate is removed): 60" long (152cm) x 33.5 " (85.1 cm) with only monitor and bracket and foot pedal holder on back side removed (33.5" (85.1 cm). If the rear lifter arm has to be removed (this is not recommended and should only be done by IA staff) and the monitor and bracket and foot pedal holder are removed, the new width dimension is 32" wide (81 cm).

vi. Dimensions for shipping (with reusable crate): 76" long x 38" wide x 88" high (193cm x 96cm x 224cm). IMPORTANT: ensure that truck has a lift gate and a door OPENING of 90 inches for clearance. An alternate, smaller footprint crate method may be used. Confirm this with IA staff.

vii. Shipping Weight with crate: approx. 668 pounds (276 kg).

viii. Bandwidth Requirements for one Scribe station: 1.5 megabits/sec of Internet connectivity 24/7 (either commercial Internet or Internet2); 1 real external IP addresses not blocked by any firewall (this is independent of the number of Scribes installed). Projects requiring more than one Scribe, please contact us for more detail on Bandwidth requirements.

B. Output Formats

i. Color images in JPEG2000 format in pixels per inch listed below. We make our jp2s via the Kakadu tools, and we control the amount of compression by specifying the -slope option (which sets the rate-distortion slope); that results in a constant image quality, rather than constant file size compression ratio. Highly compressible images - for instance, largely blank pages - therefore compress down to a very small size, because that's all that's needed to achieve the desired quality. Such pages will have a higher compression ratio than pages full of text. We don't use tiling and our jp2s have just one layer. JP2 profile as follows:

Compression - slightly lossy
Profile-icc: 3024 bytes
Description: sRGB IEC61966-2.1
Manufacturer: IEC http://www.iec.ch
Model: IEC 61966-2.1 Default RGB colour space – sRGB
Copyright: Copyright (c) 1998 Hewlett-Packard Company

ii. Optical Character Recognition (OCR) in two XML formats: ABBYY and DjVu. (ABBYY 9.0 is currently used, but may be changed in the future). As new OCR versions and alternative vendors become available, a review will be conducted. If a new version or vendor is deemed as good as or better than the existing version or vendor it may be implemented without any notice. A list of current OCR language codes is available upon request. OCR XML character format is UTF-8.

iii. XML for metadata from MARC.

iv. XML for operational metadata collected during digitization.

v. Searchable PDF/A.

vi. XML structural metadata for monographs and serials includes: pagination (when page numbers are printed on the book leaves), front/back cover, title page, copyright page and tissue paper (if found in the book.)

vii. The formats listed above will be delivered from the Internet Archive servers to the Internet via HTTP. Testing at the Library of Congress suggests that downloading an entire book file from the IA site takes approximately 90 seconds.

viii. The Library Partner may download as many copies of each Public Domain file as they wish.

C. PPI and Size

i. For Canon Mark I the PPI chart is below and shows both the book and the PPI settings. These are chosen to optimally capture a given size book. Some centers still use Mark I cameras, most have been converted to Canon Mark II.

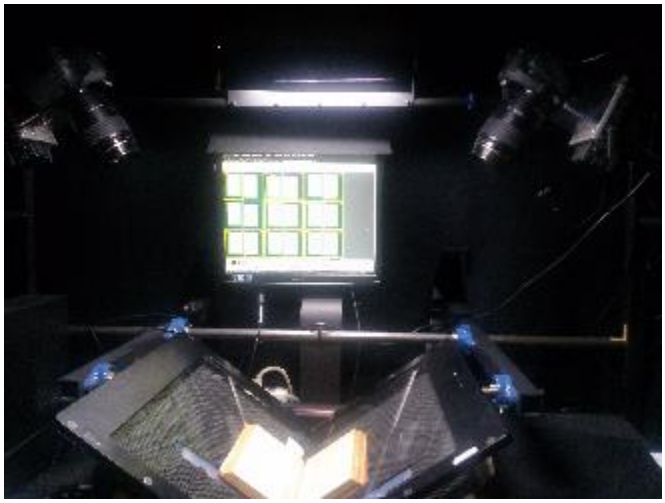
PPI	Height (inch)	Width (inch)	Height (cm)	Width (cm)	
300	16	9.25	40	23	
400	10.75	6.75	27	17	
500	8.75	5.5	22	14	

ii. For Canon Mark II's the PPI chart is below.

PPI	Height (inch)	Width (inch)	Height (cm)	Width (cm)	
650	8	5.5	20.3	14.0	
500	10.5	7.5	26.7	19.1	
350	15	10.5	38.1	19.1	

D. Image Capture

i. The Scribe machine currently captures page images with two digital single lens-reflex (DSLR) cameras, specifically the Canon model 5D, Mark II - 21 mega-pixel camera (http://is.gd/lZVLMo) and the Canon EF 100mm f 2.8-macro lens (http://is.gd/IuRG). In all regional digitization centers and most satellite digitization centers Canon model 5d, Mark ii – 21.1-mega pixel cameras are being used. IA may in the future evaluate and test newer camera models as they come onto the market to determine if they will provide similar or better performance. Note: there is a possibility that some dust particles may be captured from the glass or lens during photographic process. While IA will try to reduce and eliminate this, this might occur and IA cannot be responsible for this. There is also a possibility that on some photographs, particularly in yearbooks, a Moiré effect may occur. IA can't control this and this not part of our standard QA or process control.



Digital Cameras Shoot from Above and Capture Both Verso and Recto Simultaneously

ii. The lighting system used for book illumination consists of LED, low-energy, low heat lights. These are similar to the previous lights, which were 5000 or 3500 Kelvin, 36 degree, and 35-watt museum-grade bulbs and that provided a smooth daylight spectrum with a high color-rendering index. If future alternative lighting methods are found to provide similar or improved results, changes to the lighting system may result.

iii. Please note that since there are two independent cameras in use, there may be a detectable difference in lighting between the recto and verso images.

iv. Reference targets: color targets (such as a ColorChecker 24) are shot at the end of each book as reference tools and may be used for ICC-based color management.

v. Image transfer: images are downloaded in real time to a Scribe management/image-processing computer. This computer is also responsible for running the camera management software that operates the camera shutters. An upload process will then happen to allow for secondary processing of the images, file derivative generation and several other operations.

E. Equipment Calibration

i. Scribe station frames are calibrated and aligned before being put into use.

ii. Cameras are calibrated per Internet Archive specifications. Cameras that no longer perform within specifications are sent to the manufacturer or repaired in-house.

3. Project Setup and Shipment

A. Project Preparation and Special Handling Requirements

i. Prior to beginning a new digitization project, several forms including a Content Agreement, Partner Form and, if necessary, a CSV-style Pick List must be completed and reviewed with IA staff. Please refer to www.archive.org/details/partnerdocs for the most current documents related to setting up a new digitization relationship between the Internet Archive and a Library Partner. Based on the volume of digitization to be undertaken, a Z39.50 connection may be established between the Library Partner and IA. If it is deemed unnecessary, IA will use the Library of Congress catalog or similar catalog to locate a corresponding MARC record. Correspondingly the CSV sheet (found at www.archive.org/details/partnerdocs) may be used to create library partner initiated metadata. Contact the Site Coordinator for details.

ii. IA and the Library Partner will review the materials chosen for digitization and determine what can and cannot be included in the process. Questionable materials may be tested before being added to the digitization workflow. Selection of all materials will follow this guideline.

iii. Library preservation personnel may meet with IA to establish and agree on any special handling of materials to be digitized, as well as how to deal with rejected materials, how to indicate when material cannot be digitized, and similar workflow processes.

iv. "Special Handling" materials that are fragile, unique or not currently in a library's general circulation must not be included in the general workflow alongside more robust materials. If special books, books with unique handling needs or the equivalent are to be digitized, IA must be notified beforehand. Deviations from this process will be in writing and, where possible, all steps will be documented with visual images for reference. Again, it is important that any materials not part of a library's general collection, or requiring special handling, be identified as such by the Library Partner and be sent in a separate, distinct shipment.

B. Shipment of Materials to the Digitization Center

i. Only materials meeting IA specifications of items to be digitized (See Appendix: Part 1, Digitization Criteria) will be delivered to the IA Digitization Centers.

ii. Materials will be packed on a library cart, unless another method is agreed upon, and safely

wrapped for transport. This is to be done by the Library Partner. Any special procedures will be determined in advance.

Books are Typically Brought to the Digitization Center in Boxes or on Book Trucks

iii. A paper copy of the item Pick List (see example in appendix) is to travel with each shipment of materials delivered to a Digitization Center. In addition, a digital copy (Excel) must be sent simultaneously to the Digitization Center Coordinator.

iv. When books are received in the Digitization Center, they are first inspected. Any exceptions such as damage, items that do not fit the digitization profile or similar anomalies, are noted. The Digitization Coordinator will notify the Library Partner of the materials' safe arrival by email unless otherwise agreed to.

v. IA will also ensure that the Pick List item count matches the number of items in a shipment. If a discrepancy with the item count is discovered, the Digitization Center Coordinator will alert the Library Partner immediately and no further work will occur until this is resolved.

4. Book Digitization Workflow

A. Metadata Retrieval and Book Loading

i. The Digitization Center Book Loader will inspect each item for possible factors that will impact its ability to be digitized (See Appendix: Part 2, Section A, page 22). These items may be digitized at a later date as new processes become available and/or as different cost structures are put into place.

ii. In order to begin loading, a book ID or equivalent (e.g. a bibliographic ID) is loaded into the IA PMA Tool in order to locate a corresponding MARC record. Concomitant with this, a CSV pick list is used to create or load bibliographic data. A quality check is performed to ensure that the item in hand and the MARC record match exactly.

iii. Books with no locatable catalog record should be accompanied by a CSV document detailing the books' relevant metadata. Any items that we are unable to digitize, both due to an inability to locate adequate metadata or due to book condition issues, will receive an explanatory reject form and be returned to the Library Partner.

iv. If a Digitization Center receives a series of items that have been cataloged within one bibliographic ID but have no discernable set of volume numbering, or are part of a set that has been catalogued under multiple titles, or any other similar cataloging anomaly, IA will work with the Library Partner to determine the best method for uniquely identifying each item. This may include supplementing each record with metadata unique to each item or otherwise creating a volume numbering system for digitization purposes. IA will not delete or add any information to the description fields within the MARC record.

v. IA's PMA Tool automatically creates a unique, persistent identifier for each item and the MARC record is attached to that identifier.

vi. Each book within a collection or project may be given a color-coded flag (to assist in book tracking), which also indicates the book identifier. Any special digitization instructions are included with the book. The book is now ready for digitization and is placed in a queue for the Scribe Operators.

B. Book Digitization Process

i. Scribe operators compare the color-coded flag containing the book identifier with the actual item in hand. This ensures that the file of digital images they create will be matched up with the correct corresponding metadata.

Books with Flags

Books with Foldouts or Maps are Identified

The Glass Platen is Gently Lowered Onto the Book Before Shooting the Next Page Spread

The Scribe Operator May Reposition the Book as Necessary During Digitization Process

The Scribe Operator Turns Each Page

ii. Images are shot through anti-static, anti-glare coated glass to minimize distortion due to page curvature. This also aids in the OCR process.

iii. Images are quality checked in the republishing phase of the digitization session and are adjusted for such things as text block cropping and image de-skewing in order to ensure proper preservation and presentation. NOTE: IA uses an algorithm to set the skew on a page. It works best on pages where there is a full text box. It does not do as good a job when there is a photo, title page, chart or text block without a lot of text. IA will attempt to manually adjust the skew in the first 10 pages of a book, but will in general not correct skew elsewhere in the book unless prior discussions have happened with the Library Partner.

iv. The digital file is inspected for missing pages, crop/de-skew issues and general presentation, as part of the Quality Assurance process.

v. The completed files are then uploaded to IA's Data Center for processing.

C. Foldouts and Maps

i. Foldouts and maps of appropriate size may be digitized in full color and can be inserted seamlessly into the leaf images of any digitized item.

ii. To preserve acceptable resolution, normal foldouts should be no larger than: 30” wide x 20” high (76cm x 50cm).

iii. For items larger than 30” wide x 20” high (76cm x 50cm), a single, lower resolution reference shot of the entire foldout is taken and is supplemented by a number of higher resolution sectional images. For example, a large map may be shot as a single reference shot PLUS 3 additional shots- 1 of 3, 2 of 3 and 3 of 3 shots.

iv. The maximum size foldout allowable is 41” wide x 27” high (104cm x 70cm). (For this size, the single reference image is shot at a reduced resolution of 107 PPI.

v. NOTE: IA also has the capability to insert a client’s image into the leaf images of any digitized item. This image may be virtually any size, as long as it is an uncompressed TIFF or JPEG image. With this capability, a Library Partner may use their own camera to capture a map or foldout at a very high resolution. IA can then insert this image into the series of images that comprises the digitized book.

Camera Height	PPI	Width & length (up to)
21.3 in. (54.1cm)	248	18" x 12" (46cm x 30cm)
24.2 in. (61.5cm)	208	21" x 14" (53cm x 36cm)
27.4 in. (69.6cm)	180	24" x 16" (70cm x 41cm)
30 in. (76.2cm)	146	30" x 20" (76cm x 50cm)
36 in. (91.4cm)	121	36" x 24" (91cm x 70cm) – also shot in sections
41 in (104.1cm)	107	41”x 27” (104cm x 70cm) – also shot in sections

Foldouts are Digitized Using an Overhead Camera After Book Digitization at the Scribe

Images are Flattened Using Magnets and are Positioned to Minimize Curvature. Multiple Shots May Be Taken to Increase the PPI if the Image is Large.

D. Other Capabilities

(Please contact IA for more details on pricing and specs)

i. Microform - IA can accommodate standard 35mm and 16mm microfilm and standard microfiche. Microfilm and microfiche digitization is presently performed at the San Francisco, California facility. Images are captured as gray scale, turned into JPG 2000 images and then processed as books. Pricing and specs are available upon request.

ii. Folios - IA is able to digitize bound, single-sided, folio-sized books, an example being a large art folio. In order to retain acceptable resolution, the ideal maximum size for a folio would be 12" wide x 18" high (30cm x 46cm) and a thickness of about 150 pages. A test digitization of potential material is recommended. Pricing and examples are available upon request.

iii. Archival Collections - IA is able to accommodate special digitization needs, such as those arising from collections of papers or ephemera. We can work with the special needs of a collection and ask only that the materials are flat and do not contain staples or clasps that must be removed in order to digitize. The Library Partner should perform any collation or organization necessary.

iv. Photographic materials - IA is able to accommodate special digitization needs, such as those arising from yearbooks, photo albums and loose photos.

v. Loose-leaf materials - Mass digitization of single sheet documents can be done either at the scribe work station or using a sheet-fed scanner; please contact IA for details.

vi. Video, Audio and LP conversion - IA is working to accommodate special digitization needs, such as those arising from audio and video materials.

vii. Yearbooks, full text bleeds and large text blocks – Note, that when images bleed to the edge, text blocks are close to the edge of the page or marginalia runs close to the edge of the page, "over-cropping" may be employed. Please consult IA for details. The following information will be sent out to Library Partners if we receive yearbooks to digitize.

For the Library Partners:

Libraries and various institutions have been very keen to have the Internet Archive digitize yearbooks. We have, over the course of our operations, digitized thousands of yearbooks over the past few years. In short, we love yearbooks!

We have learned that there are four areas where the physical design, size, binding or paper type of a yearbook (or in some cases 'text-blocks") might not yield an online image that meets the expectations of the content provider or end user. These four areas are detailed below:

1. Glare appears on portions of the pages. The cause of this issue is black and/or glossy physical pages. While we have tried various approaches to solve this problem, we do not have a solution that is 100% effective.

2. Center text (captions that cross from one page to the other) or images are cut off. Many yearbooks have text or images that run into the center gutter. Due to our non-destructive digitization methodology, we can't capture images or text that runs into the gutter. The only partial solution we have is to over crop the center gutter, which does not guarantee the capture of text or images that are deep in the gutter. A decision needs to be made before imaging begins about pre-approving center over crop on all pages or no pages of an entire shipment of yearbooks.

3. Text or images that run to the outer edge of the page may get cut off. To address this situation, we will, as a standard procedure, over crop the outer pages. This will result in up to a 0.5-inch dark colored border around each of the pages.

4. Moiré pattern appears on images. On occasion, the printing of an image or picture may result in a digitized image that displays undesired digital artifacts in pictures or photographs on a page. This is due to a rare interference pattern that occurs when the grid of the printed image is overlaid with the grid of the camera sensors, and is more common when digitizing halftone images. The results can range from a mild-to-extreme moiré pattern. We can't predict when this will happen and at this stage do not have a solution for this condition.

We are working to improve our capabilities to address these situations, but until a better solution is found, please be aware that digitization of these types of materials could yield the aforementioned results. As such, we will not be able to re-image or correct items that display these conditions.

4. Post Digitization Processing

A. Image Processing

Uploaded images are processed to create storage and access files. (See Section 1, Part B, page 4 for a more general list of output files).

i. Detailed list of files
   1. ID.pdf
   2. ID_jp2.zip
      - zipped folder of the book
      - [ID]_nnnn.jp2
   3. ID_meta.mrc
   4. ID_meta.xml
   5. ID_metasource.xml
   6. ID_raw_jp2.zip - unprocessed storage format (no bookplate)
   7. Scandata.zip

i. Metadata will reside in meta.xml file, and may include the following
   1. Identifier
   2. Identifier/Bib ID (IA identifier and local ID from Pick List)
   3. Contributor
   4. Title
   5. Volume
   6. Creator (if in MARC record)
   7. Publisher (if in MARC record)
   8. Collection/s
   9. Operator
   10. Scanner (machine)
   11. Scandate
   12. Identifier/Access (URL for accessing this book is found by using www.archive.org/details/identifer)

B. Background Processing

i. Initially the digitized image is captured as a " raw" JPG. This JPG is appx 3-5 MB and is run through a JPEG2000 conversion to generate a raw JPEG2000 for storage. The raw JPEG2000 is then turned into a processed master, which is used to generate the access formats.

ii. Storage format – raw JPEG2000 is a compressed, lossy, un-cropped, non-rotated, non-de-skewed, non-light compensated JPEG2000 file. There is a contrast enhancement process step that is done here. It expands the color-value range of 30-240 to 0-255. This means that the darkest -12% of the total range gets flattened to black and the lightest -6% gets flattened to white, with everything else in-between stretched accordingly. Image sizes vary depending on the complexity of the page, but are typically in the 0.5 MB range, yielding an approximate compression ratio of about 10:1 relative to the raw JPG (JPG is approximately 5 MB/image.)

iii. Processed master – lossy, cropped, rotated, de-skewed, light compensated JPEG2000. Image sizes may vary depending on complexity of the page, but are typically in the 0.5 MB range, yielding an approximate compression ratio of 10:1 relative to the camera raw image (JPG is approximately 5 MB/image).

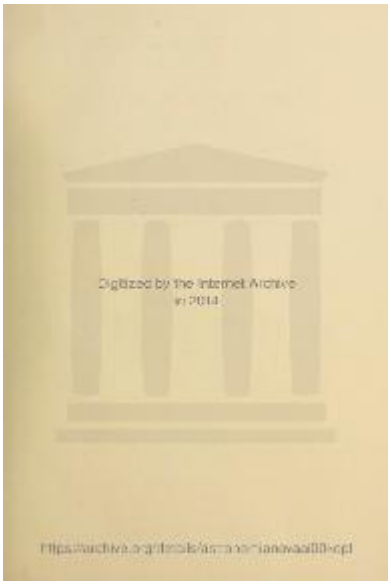
iv. Access format – the processed JPEG2000 masters are compressed in a JPEG2000 format, which feeds into the OCR and book generation tools. Image sizes may vary depending on the complexity of the page, but are typically in the 0.5 MB range, yielding an approximate compression ratio of 10:1 relative to the initial JPG (JPG is approximately 10 MB/image). Also ReadOnline view and PDF both of which are OCR'd.

v. Note: all compression ratios might vary based on which version of ABBYY is used and on specific software parameters. These numbers are for reference purposes only.

vi. Quality settings will vary based on vendor tools used. For example, a quality setting of 50 on a scale of 1-100 was used for the LuraTech (or equivalent) PDF compressor. This setting was determined based upon user surveys. As improved software becomes available, vendor selection may change.

C. Internet Archive Book Plate

i. A bookplate may be digitally inserted in the beginning pages of each digitized item. The bookplate contains the attribution "Digitized by the Internet Archive", the year digitized, and the URL for the item. The bookplate algorithm will find a blank image or a near blank image within the first 10 pages of the book. This algorithm is fairly conservative, so if it does not find a sufficiently blank page, no bookplate will be inserted. This is to avoid obscuring text on a page. It is rare that an item does not have at least one blank page within the front matter, so it is rather infrequently that we will dispense with the bookplate insertion altogether.



Internet Archive Bookplate

D. Completion of Book Processing

i. For digitization performed on-site in a partner library, the typical turnaround of a book cart is 72 hours, from arrival to return. A Digitization Center that consists of 10 Scribes running 1 shift will complete approximately 500 monographs per week. The Internet Archive process is scalable and can be made larger to handle higher weekly capacity or be as small as 1 Scribe. For shipments that will be sent to a regional center, turnaround time will be estimated prior to shipping of the materials.

ii. IA's goal is to derive and upload books for web access within 24-48 hours after digitization.

iii. An internal Quality Assurance process is performed inside the Digitization Center after the books are made available online. (The Quality Assurance process is outlined in Section 5).

iv. Items that pass the Quality Assurance process are ready to be "checked out" and returned to the Library Partner. Return shipping details shall be established prior to the commencement of the digitization process. Any rejected materials are also returned at this time.

v. At month's end, a final curation by IA staff is undertaken and a bill or invoice is issued. Materials will be available online prior to this and may be downloaded, but until a final invoice is issued and the curation occurs, there is a chance that minor changes to metadata or the files may still occur.

E. Check-Out Process and Material Return

i. The Digitization Center staff performs a final count of items to ensure all Library Partner material is included.

ii. The Digitization Center staff packs the books onto library book carts, shipping containers or other such receptacles per the guidelines established between the Library Partner and IA.

iii. The Digitization Center Coordinator may generate an updated Pick List that now contains corresponding IA book identifiers and indications of any rejected materials.

iv. The books are then shipped or returned to the Library Partner.

5. Quality Assurance (QA) Process

A. PPM

PPM is a measure of how many pages per million could have a quality error. The formula used to calculate ppm is the total number of pages rejected (a) divided by total number of pages QA'd (b), multiplied by 1,000,000 ((a/b) x 1,000,000).

In order to illustrate this methodology, let's consider a hypothetical bin size of 138 books. For this bin size, Internet Archive would QA 20 books. Let's say one book had a couple of pages with cropped text. According to our prior QA methodology, 1 rejected book out of 20 would generate a passing bin. However, under the new ppm measurement, we would consider how many pages were QA'd (ex: 5,647 pages) vs. how many pages were rejected (one page of cropped text / missing page / blurry page causes the whole book to be rejected.) So if this bin had 238 pages that were rejected for any major code, the ppm for this bin would be (238 / 5647) x 1,000,000, or 42,146 ppm. This is outside (above) our current acceptable ppm measurement of 20,000 ppm, so the bin would fail.

B. Three Major Phases of the QA Process

i. At a Republishing Station - Before any digitized images are uploaded, the republisher will quality-inspect each image. The republisher will review the images for missing spreads, crop/de-skew issues, accurate page labelling (i.e. title page, covers, TOC) and will add any notes regarding defects in the book to the file (i.e. missing pages, tight binding, torn pages.)

ii. After the Images are Online – A PPM-based Quality Assurance process.

iii. After the Invoice is Sent - Errors brought to IA’s attention will be reviewed and dealt with in a timely manner. IA and the Library Partner will decide if any materials need to be re-digitized or if the problems can be resolved within the digital file, post-derive. Re-imaging is avoided when possible, as it is usually the most expensive solution.

C. QA Process in the Digitization Centers

i. Each day the Digitization Center will review a set of books from the previous two business day's digitization production. The number of books to be reviewed will depend upon the total number of books in the set.

Books in set	2-8	9-15	16-25	26-50	51-90	91-150	151-280	281-1200
Number to QA	2	3	5	8	13	20	32	99

ii. The QA technician will select a representative sampling, which reflects a broad combination of operators and machines and conforms to the statistical chart.

iii. The online digital books are then inspected using specific criteria. (See Appendix: Part 2, Section D,). If found, errors or defects are noted and added to an IA tracking form.

iv. Example: If 125 books have been digitized in the previous two business day's production, bin 6 will be selected and 20 books will be 100% inspected.. The following steps are then followed:

1. If a "fail" is generated, the Digitization Center Coordinator will review the errors or defects to ascertain if they were generated from outside the Digitization Center or from within. Engineering is notified of any problems out of the Digitization Centers' control (i.e. a missing access file) and Scribe operators are notified if any problems are directly a result of their work (i.e. missing page spreads).

2. If an error is generated from within the Center, the Digitization Center Coordinator will follow a pre-determined set of process steps, ultimately culminating in a recommendation to approve the lot, or a portion of the lot, with appropriate corrective actions identified. At this stage the Director of Books or the Quality Assurance Librarian may be involved and would approve a deviation. A corrective action report will be generated for rejected lots. This will be reviewed with engineering and operations management for longer-term solutions or corrective action.

(Please refer to Appendix: Part 2, Section B for a more detailed list of QA codes, their definitions and how they are used.)

6. Digitization Center Locations

Materials will be digitized at one of 8 regional Internet Archive Digitization Centers: San Francisco, California; Ft Wayne, Indiana; Beltsville, Maryland; Boston, Massachusetts; Princeton, New Jersey; Toronto, Canada; Shenzhen, China; and London, UK.

There are also 20+ smaller digitization centers located in Brasilia, California, the District of Columbia, Idaho, Illinois, North Carolina, Hawaii, Utah, China, Alberta, Canada, South Africa, Guatemala and the United Kingdom.

Below are pictures of several of the centers:

Internet Archive Satellite Digitization Center at the John Carter Brown Library- Providence, Rhode Island

Scribes at the University of Toronto Digitization Center in Toronto, Canada

Scribes at the Wellcome Library Digitization Center in London, UK

Scribes at the Allen County Public Library Digitization Center in Fort Wayne, Indiana

Internet Archive Digitization Center at the National Agricultural Library in Beltsville,

7. Internet Archive Contact Information

For General Inquiries, please contact digitallibraries@archive.org

Name	Role	Phone	Email
Jude Coelho Process Manager	Problem resolution, engineering and tech support	415-561-6767	judec@archive.org
Digitization Center Coordinator (See list below)	Daily Operations Ongoing partner support		(See list below)
Ken LeTran Systems Engineer	Engineering and systems support		ken@archive.org
Shelia DeRoche	Quality	202-707-6988	shelia@archive.org
Andrea Mills	Partner Specialist, Digital Libraries	416-946-7488	andrea@archive.org
Gemma Waterston Batson	Brand Manager, Digital Libraries		gemma@archive.org

Digitization Center Coordinator Contact Info

San Francisco, CA	Jesse Bell	bell@archive.org
Fort Wayne, IN	Jeff Sharpe	jeffs@archive.org
Boston, MA	Tim Bigelow	Tim.B@archive.org
Beltsville, MD	Shelia DeRoche	shelia@archive.org
Princeton, NJ	Stacy Argondizzo	stacy@archive.org
Toronto, Ontario	Gabe Juszel	gabe@archive.org
Euston, London	Chris Booth	cbooth@archive.org
Satellites	Elizabeth MacLeod	emacleod@archive.org

Appendix

1. Digitization Criteria

Criteria used to determine if materials may be digitized are listed below and will include, but not necessarily be limited to, the following:

A. Book Condition / Preservation Standards

i. Library preservation personnel will meet with IA to establish and agree on any special handling of materials to be digitized. If special books, rare books or the equivalent are to be digitized, IA must be notified beforehand if special handling processes are to be used. TO REPEAT: IA MUST BE NOTIFIED IF A BOOK IS CONSIDERED NOT TO BE A GENERAL CIRCULATION BOOK OR IS CONSIDERED SPECIAL, FRAGILE OR REQUIRES EXTRA CARE.

ii. Non-circulation materials that are fragile, unique or not currently in library's general circulation must not be included in the general IA digitization workflow. It is important that any materials that require special handling, be identified as such by the Library Partner and be sent in a separate, distinct shipment.

iii. Books with covers that are separating or that otherwise appear fragile, may be rejected unless a decision to the contrary, prior to digitization, has been made between IA and the Library Partner.

B. Materials with Multiple Titles / Multiple Volumes

i. If a Digitization Center receives a series of items that have been cataloged within one bibliographic ID but have no discernable set of volume numbering, or are part of a set that has been catalogued under multiple titles, or any other similar cataloging anomaly, IA will work with the Library Partner to determine the best method for uniquely identifying each item. This may include supplementing each record with metadata unique to each item or otherwise creating a volume numbering system for digitization purposes. IA will not delete or add any information to the description fields within the MARC record.

ii. A Partner Meta app with CSV forms is used to capture metadata. https://archive.org/details/PartnerMetaApp

C. Size, Shape, and Pricing Requirements

(Materials not fitting the requirements shown below may be returned, un-digitized.)

i. 9.5” wide x 14.5” high (24cm x 36.8cm) maximum, 3” wide x 3” high (7.6cm x 7.6cm) minimum.

ii. Items thicker than 3” (7.6cm) will be reviewed in order to determine if the book cradle can accommodate them and to ensure the glass platen can reach into the gutters.

iii. On average, books should be approximately 200 pages or greater. If a collection consists mainly of items with fewer than 100 pages, IA and the Library Partner will review in order to ensure that the quoted price per digitized page can be maintained.

iv. IA has the capacity to digitize foldouts and maps. (See Section 3, Part C) Testing and decisions about specifications should occur prior to digitization. Foldouts and maps will have a lower resolution, or PPI, than regular image captures. Foldouts and maps have a higher price point than regular image captures.

v. General pricing details may be found here and are subject to change - https://archive.org/details/iacontentscanningform

D. Book Style

i. Ideal candidates for digitization are monographs and serials – that open from left-to-right or right-to-left. Books that flip ‘up’ to open may not be able to be digitized on a Scribe, but might be on the foldout machine.

ii. Soft cover books are acceptable as long as they are bound.

E. Paper Style / Print Quality

i. Most paper styles may be digitized except highly acidic paper that has already degraded and is disintegrating. Please note that if any fragile paper is to be digitized, IA will conduct a review of the amount of extra time needed for digitizing these materials.

ii. Pages should not be excessively dusty or have excessive mildew or mold.

iii. All pages should have the bolts pre-cut. Unless otherwise instructed, books with bolts (uncut pages) will not be digitized unless arrangements are made with the Library partner.

iv. Scribe operators should be able to lift and turn the pages with normal effort. Pages with adhesive or that are otherwise tacky will not be digitized.

v. Microfilmed reproductions should be reviewed with the Digitization Center before being imaged. If the resulting text resembles a film negative or photocopy, we will not digitize it. If microfilm reproductions have more than one page on each leaf, these will also be rejected. Any deviation must be agreed to by IA and the Library Partner.

F. Gutters / Margins

i. All books, rebound books in particular, will be reviewed to check for tight gutters and narrow margins along all sides of the text block.

ii. Tight bindings that will not lay open for digitization per IA specification limits may be rejected.

iii. If the page text is less than a quarter inch away from the gutter, the book will not be digitized, unless agreed to with the Library Partner and IA.

iv. Text that runs to the edge of the page or margin may be digitized, but the final presentation may be less than optimal. The decision to digitize material of this nature will be made prior to digitization.

G. Bibliographic Metadata

i. When two or more titles are bound within the same volume (bound-widths) each item will need to be designated with its own pull slip and library identifier (i.e. Bib ID). The Library Partner will prepare this. Each title will be considered a separate book and will be digitized as such. IA and the Library Partner must discuss any deviations from this rule.

ii. Books that are post 1923 may be digitized, but should be discussed with IA and the Library Partner prior to digitization.

iii. If IA is unable to locate an item's MARC record, the book will be returned to the Library Partner pending further information. For items without MARC records, a csv of the items' metadata can be provided by the Library Partner. See Part 3, Section A for more information.

H. Archival Materials

The Internet Archive has extensive experience in this area. Please contact your regional Coordinator or robert@archive.org.

2. Rejection / Error Codes and Resolution

A. Rejection Codes

Any items that we are unable to digitize, due to an inability to locate an adequate MARC record or due to condition issues, may receive an explanatory reject form and be returned to the Library Partner. The following issues may be identified on the form.

- Item is fragile or has no binding
- Item record is unreadable
- Item is damaged
- Item is still in copyright
- Item is an exact duplicate of another item on list
- Item has foldouts that are too large for imaging
- Item is not a book
- Item is outside the language parameters
- Item is too large for imaging
- Unsuccessful link to metadata
- Pick List error
- Items margins/gutters are too tight for the Scribe
- Item is missing more than five pages
- Item has an unacceptable number of volumes bound together
- Not available
- Item is not on shelf (missing/lost)
- Item is not on shelf (checked out)
- Pagination issues (section[s] bound out of order or upside down)
- Item has brittle pages, deemed unsuitable for digitization
- Item has skewed text (to the point of being unreadable)
- Item is too small to be imaged
- Item requires special handling
- Item has more than five uncut pages (pages must be separated)
- Vellum
- Withdrawn

B. Error Codes – Definitions and Resolutions

IA uses a variety of codes to define and describe errors that may be found in digitized items. Error resolution falls into three forms of correction: post-derive, re-imaging or a consultation with the Library Partner. If IA is unable to correct problems using any of these methods, then the book is rejected and re-imaged if possible. In addition to the thorough in-house QA performed within the Digitization Centers, if the Library Partner discovers any problems with an online item, IA will attempt to resolve the error within 30 days of being notified.

i. FREEZE CODES, part A

Items receiving these codes will be re-imaged using the same identifier and URL. If we are unable to re-image or otherwise correct any problems post-derive, the Library Partner will not be billed for this item.

- Uploading Issues

  - 111 Book uploaded from scribe before completed -- incomplete item.
  - 112 Missing / corrupted files.
  - 114 Cameras assigned incorrectly.

Resolution: Material is re-imaged.

- Metadata

  - 123 Possibly not in public domain.
  - 124 Removed by request of copyright holder or library.

Resolution: If material is in copyright, the item is removed from the IA search engine. If material is in question, the Library Partner is consulted and appropriate action is taken.

- Images

  - 130 Cropped text.
  - 131 Blurred page(s).
  - 132 Missing page(s).
  - 133 Front/Back cover missing.
  - 134 White streak in images that obscures text.
  - 135 Book was digitized twice; this copy to be darkened
  - 136 Text is washed out or overly dark. This should be used when the lighting is so bad that it affects human readability and/or OCR-ability.

- 137 Object (fingers, shadows, string, paper, etc.) obscuring the text.
- 138 Glass not centered in gutter; text is distorted or cropped.
- 139 Foldout shot as a normal page and checked in.

Resolution: Material is re-imaged.

ii. FREEZE CODES, part B

These codes are used for books that have fixable problems, but are not yet in billable condition.

- 140 Book and metadata do not match.
- 142 Tissue pages marked incorrectly.
- 143 Anomaly in image format.
- 144 Left/right pages are reversed.
- 145 A R-L book was imaged L-R

Resolution: For items receiving codes 140-143 a post-derive correction is attempted. For items receiving codes 144-145, if a post-derive correction doesn't fix the problem, the material is re-imaged.

iii. INFORMATIONAL CODES

- 151 Bookplate or watermark missing or corrupt.
- 152 Copyright evidence was reported incorrectly.
- 156 Dust/debris spots on lens visible in images.
- 159 Color shift; cameras improperly white balanced.
- 160 Light/dark pages (intermittent).
- 161 Light/dark pages (throughout).
- 162 Pages skewed.
- 163 Color cards show in access formats.
- 166 Image of cradle is visible at front or back.
- 167 Different crop-box sizes in same spread.
- 168 Bad crop at page edges/gutter.
- 169 Duplicate page spreads imaged.
- 170 Page types not marked or marked incorrectly.
- 173 Page numbers not marked or marked incorrectly.

Resolution: These codes are informational only and do not affect the readability of the item. Books with these problems can be approved, but the errors should be noted for purposes of re-training, equipment repair or calibration, etc. Variations from this will be discussed in advance between IA and the Library Partner.

C. Re-Imaging Process

i. For materials to be re-imaged, a request for the re-delivery of those items is submitted to the Library Partner approximately once a month. Materials are re-imaged using the existing URL.

3. Post-Digitization Reporting Tools

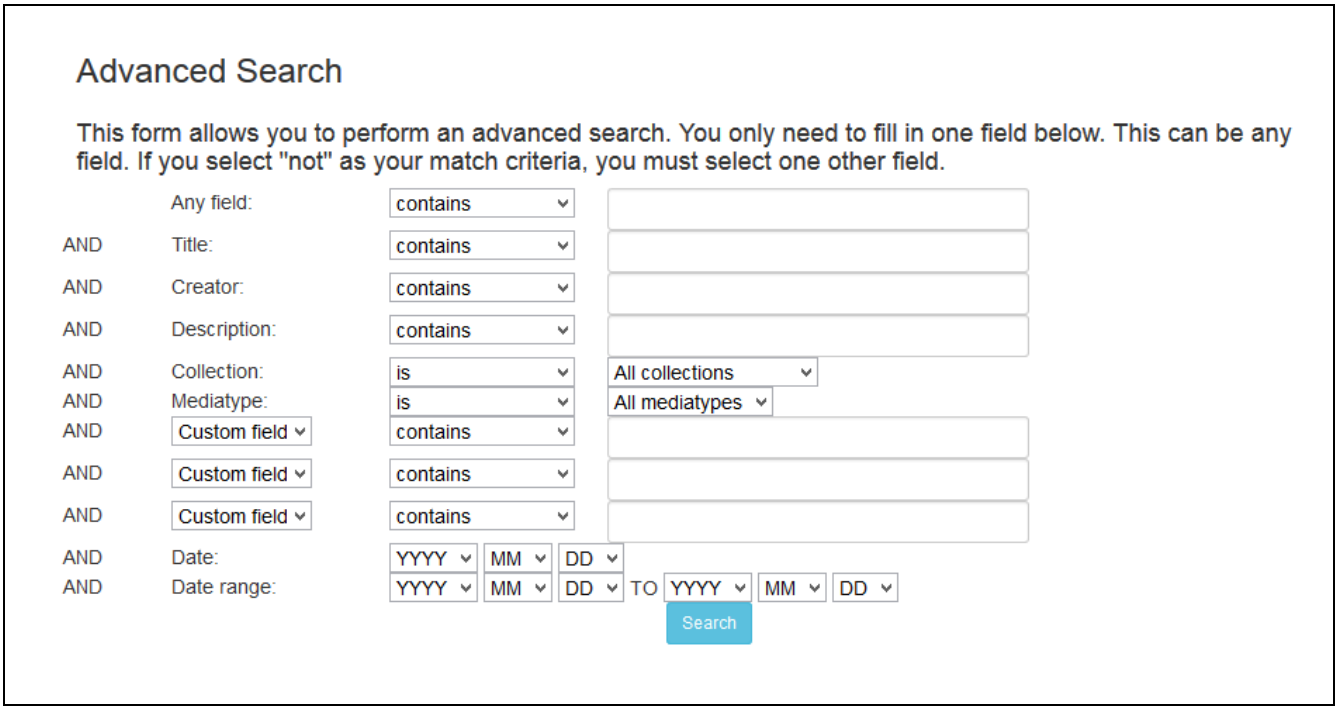
All Library Partners have access to the IA Advanced Search Engine, found at: http://www.archive.org/advancedsearch.php. This is a helpful reporting tool that may be used to search and review books that have been digitized, uploaded, QAed and curated.

Fields that are viewable in the Advanced Search Engine include:

- title
- creator
- collection
- contributor
- sponsor
- image count
- public date

A. The Advanced Search Engine

The Advanced Search form is relatively simple to operate. Users only need to enter one search term into any search field in order to produce results. Note: If you select "not" as your match criteria, you must select one other field. https://archive.org/advancedsearch.php



B. The Advanced XML Search

This is used similarly but has some significant differences. This search tool is also located at: http://www.archive.org/advancedsearch.php. An example of how it is used follows. Please refer to the "Help with CSV and Excel" section in the previous URL for tips on searching and information about known bugs.

Example: If a Library Partner wished to see how many pages were digitized in a given month, August 2008 for instance, they would perform the following search:

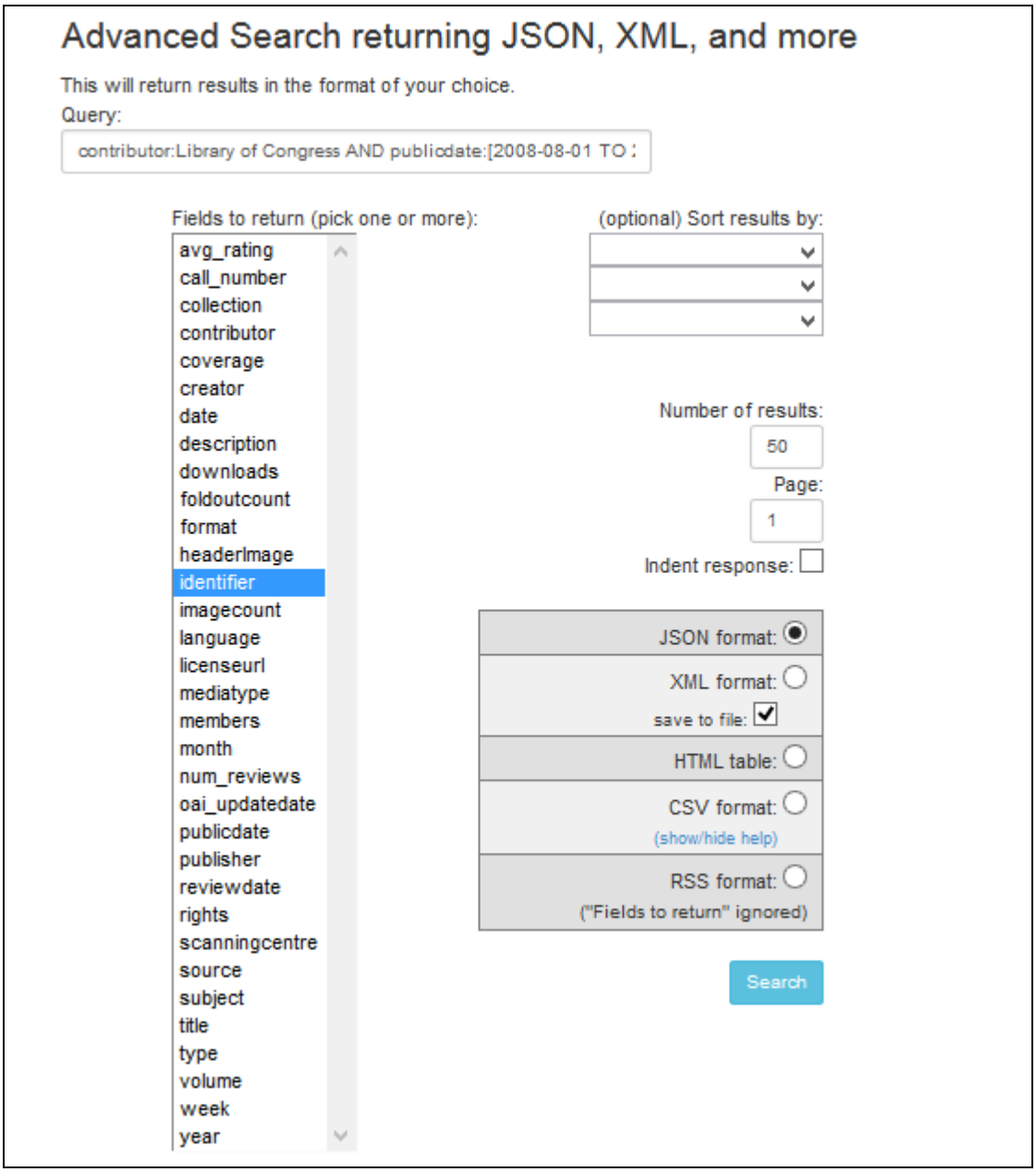
i. In the Advance XML Search "Query" field, type:
contributor:(library of congress) AND publicdate:[2008-08-01 TO 2008-08-30]
Important: *Parenthesis "()" must be around the contributor name.*
No spaces after the colon ":".

ii. Then, by holding down the "Shift" Key, highlight the fields you want to export to Excel. For example: "date, identifier, imagecount and title"

iii. Click the radio button for "CSV" and click "Search". There might be a slight delay as the search is executed. In this query, approximately 2,143 results will be returned.

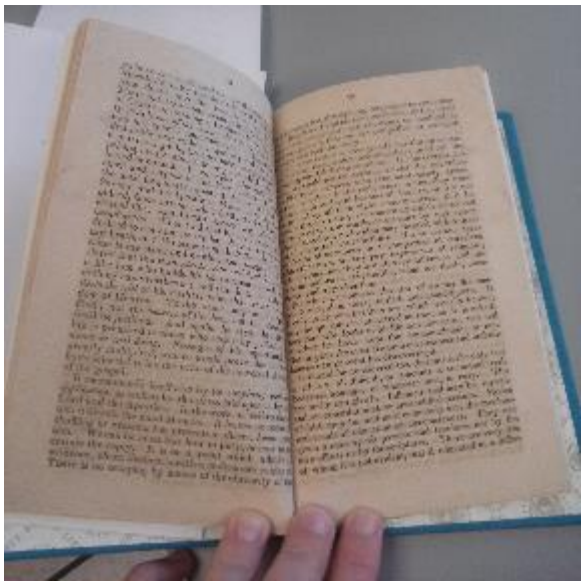
iv. When prompted, save the ".csv" file to your hard drive and note the location saved.

If you open this document in Excel, the data will be parsed into columns so that you may sort or otherwise manipulate the data.

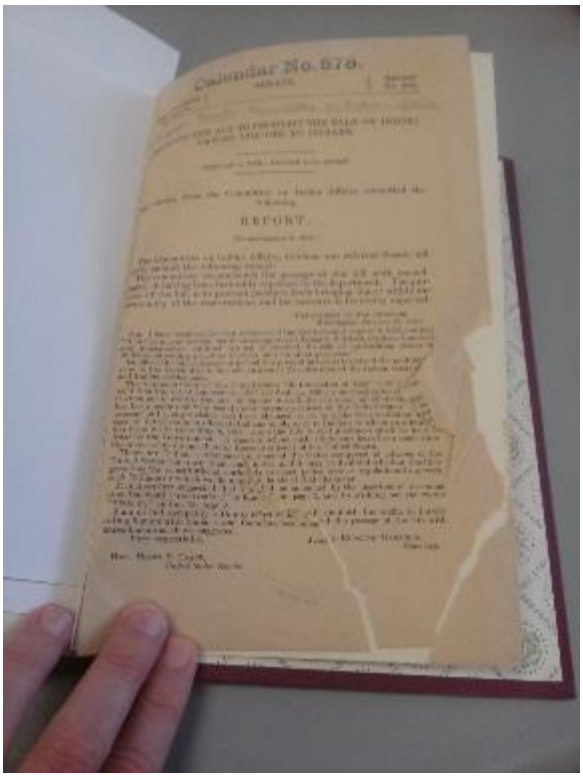


4. Examples of "Problem" Books

Uncut / Bolted Pages



Tight Gutters / No Inner Margin



Pages Torn / Damaged